擬似LiDAR表現の再考

Xinzhu Ma[1], Shinan Liu[2], Zhiyi Xia[3], Hongwen Zhang[4],
Xingyu Zeng[2], and Wanli Ouyang[1]

[1] The University of Sydney, SenseTime Computer Vision Research Group, Australia
{xima0693@uni., wanli.ouyang@}sydney.edu.au
[2] SenseTime Research, China
{liushinan, zengxingyu}@sensetime.com
[3] Dalian University of Technology, China
xiazhiyi99@mail.dlut.edu.cn
[4] Institute of Automation, Chinese Academy of Sciences, China
hongwen.zhang@cripac.ia.ac.cn

概要最近提案された擬似LiDARベースの3D検出器は、単眼/ステレオ3D検出タスクのベンチマークを大幅に改善する。しかし、その根本的なメカニズムは、研究コミュニティにとって不明瞭なままである。本論文では、擬似LiDAR表現の有効性を詳細な調査を行い、データ表現そのものではなく、座標変換に由来することを観察する。この観察に基づき、我々はPatchNetと名付けた画像ベースのCNN検出器を設計し、より一般化され、擬似LiDARベースの3D検出器としてインスタンス化することができる。さらに、我々のPatchNetの擬似LiDARデータは画像表現として構成されているため、既存の2D CNN設計は入力データから深い特徴を抽出し、3D検出性能を高めるために容易に利用することができる。我々は、提案するPatchNetが既存の擬似LiDARベースの対応する全てのデータセットを凌駕する、挑戦的なKITTIデータセットで広範な実験を行う。コードはhttps://github.com/xinzhuma/patchnetで公開されている。

キーワード:画像ベース3D検出、データ表現、画像、pseudoLiDAR、座標変換

1 Introduction

3次元物体検出は、自律走行やロボット工学など様々な分野で広く応用されているため、産学双方から注目されている。既存のアルゴリズムは、周辺環境の正確な3D点群を提供するLiDARセンサーに大きく依存している。これらのアプローチは素晴らしい性能を達成しているが、高価な機器への過度の依存により、その応用の見込みは制限されている。
急速に発展するLiDARベースのアルゴリズムと比較して、RGB画像のみから生成される3D検出[7,6,20]の結果はかなり遅れている。これは、この問題の非定型的な性質に起因しており、未観測の深さ次元に関する明示的な知識の欠如が、タスクの複雑さを著しく増大させる。直感的な解決策は、畳み込みニューラルネットワーク

(CNN)を用いて深度マップを予測し[1,10,13]、利用可能な深度情報を持たない場合はそれを用いて入力データを補強する。推定された深度マップは3Dシーン理解に役立つが、それによってもたらされる性能向上はまだ限定的である。

最近提案されたいくつかのアルゴリズム[35,24,36]は、推定された深度マップを擬似LiDAR表現に変換し、変換されたデータにLiDARベースの手法を適用する。驚くべきことに、このシンプルかつ効果的な方法は、難易度の高いKITTIデータセットにおいて、検出精度の大幅な向上を達成した。しかし、なぜそのような表現がこれほどまでに性能向上をもたらすのかは不明である。支持者の経験的説明によれば、表現の選択は3D検出システムの重要な成功要因である。画像表現と比較して、擬似LiDARはオブジェクトの3D構造を記述するのに適しており、これが性能向上の主な理由であると考えている。しかし、直接的な証拠がない場合、この記述の正しさはまだ疑問の余地がある。

本稿では、この現象の本質的な理由を探ることを目的とする。具体的には、先行研究に基づき、入力データの表現を除き、擬似LiDAR[35]と同等の実装であるPatchNet-vanillaと名付けられた画像表現ベースの検出器を注意深く構築する。この検出器により、これら2種類の表現が3D検出タスクに与える影響を深く比較することができる。他の研究[35,24,36]の議論とは異なり、PatchNet-vanillaとpseudo-LiDAR[35]の性能は完全に一致しており、データ表現が3D検出性能に影響を与えないことがわかる。さらに、入力データに対してアブレーション研究を行い、実際に重要なのは、画像座標系からLiDAR座標系への座標変換であり、カメラのキャリブレーション情報を暗黙のうちに入力データにエンコードしていることを観察した。

PatchNet-vanillaはまた、擬似LiDAR表現が画像ベースの3D検出の精度を向上させるために必要でないことを示唆している。生成された3次元座標を入力データの追加チャンネルとして統合することで、我々の3次元検出器は有望な性能を得ることができる。さらに重要なことは、このアプローチは他の画像ベースの検出器にも容易に一般化できることである。また、一種の非グリッド構造化データとして、擬似LiDAR信号は一般的に処理にポイントワイズCNN [29,30]を必要とすることに注意してください。しかし、これらの技術の開発は、標準的なCNNにはまだ遅れをとっている。この観点から、画像ベースの検出器は、擬似LiDARに基づく検出器を凌駕するはずである。この仮説を検証するために、我々のオリジナルモデルを拡張したPatchNetが提案され(例えば、より強力なバックボーンネットワーク[15,16]を使用)、KITTIデータセットにおいて他の擬似LiDARベースの検出器を凌駕した。さらに、エンドツーエンドの3D検出器を訓練できるようにするなど、画像を直接ネットワークの入力として使用することによる利点もある。以上の理由から、画像表現に基づく3D検出器は、より大きな開発可能性を持っていると主張する。

要約すると、本論文の貢献は以下の通りである: 第一に、十分な実験的実証を通じて、擬似LiDAR表現が有効である理由は、データ表現そのものではなく、座標系変換であることを確認した。

第二に、擬似LiDAR表現は検出性能を向上させるために必要ではないことがわかった。空間座標を統合した後、画像表現に基づくアルゴリズムも、同じ性能を上回らないまでも、競争力のある性能を達成することができる。第三に、より強力な画像ベースの深層学習技術のおかげで、最先端の性能を達成し、画像表現ベースの3D検出器の可能性を示す。

2 関連研究

2.1 画像表現に基づく3D検出器

この範囲の初期の研究のほとんどは、2D検出器と同じパラダイムを共有している[12,32,40,9,22,21,14]。しかし、物体中心の3次元座標(x, y, z)の推定は、画像の外観のみから絶対的な物理位置を特定する曖昧さがあるため、より複雑である。Mono3D[6]は、事前知識(例えば、オブジェクトサイズ、接地面)を用いた3Dオブジェクト提案生成に焦点を当てている。Deep3DBox[26]は、3Dバウンディングボックスが2D検出バウンディングボックスに密接に適合するという事実に基づいて、幾何学的制約を導入している。DeepMANTA[4]は、車両がよく知られた形状を持つ剛体であるため、キーポイントを用いて3D車両情報を符号化する。そして、DeepMANTAにおける車両認識をキーポイント検出とみなすことができる。ROI-10D[25]の拡張段階は、追加の深度推定器[10,5]によって提供される深度情報を利用し、それ自体が自己教師付きで学習される。Multi-Fusion[38]では、2Dボックス提案生成とそのネットワークの3D予測部分の両方について、事前に訓練されたモジュールからの視差推定結果を利用するために、マルチレベルフュージョンアプローチが提案されている。MonoGRNet[31]は、漸進的な3Dローカライゼーションと、意味的な手がかりのみに基づく3D情報の直接学習のための4つのサブネットワークから構成される。MonoDIS[34]は、2Dと3D検出のための損失を分離し、これら2つのタスクをエンドツーエンドで共同学習する。M3D-RPN[2]は、画像表現を入力とする現在の最先端技術であり、非共有重みの複数の2D畳み込みを使用して、2Dと3Dボックスの共同予測のための場所固有の特徴を学習する。上記のアプローチは、様々な事前知識、事前学習モデル、より強力なCNN設計を利用しているが、性能を向上させるために擬似LiDARデータを利用しようとはしていない。我々の研究は、擬似LiDARデータから有用な情報を抽出することで、画像ベースの手法の検出精度を向上させることを目的としており、これはこれらのアプローチを補完するものである。

2.2 擬似LiDAR表現に基づく3D検出器

近年、いくつかのアプローチ[24,35,36,39]が単眼3D検出タスクの性能を大幅に向上させている。共通しているのは、まず入力RGB画像から深度マップを推定し、カメラのキャリブレーション情報を活用して擬似LiDAR(点群)に変換することである。

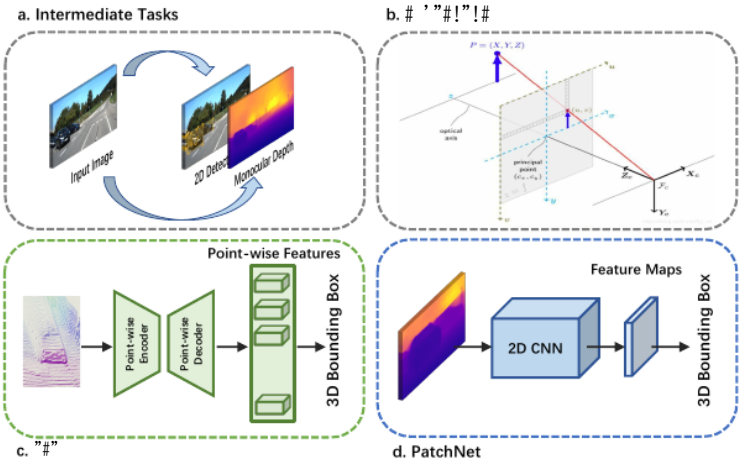

図1. 擬似LiDARベースの手法[24,35,36]とPatchNetの比較。両者とも既製のモデルを用いて中間タスクを生成し(a)、画像座標をワールド座標に投影する(b)。擬似LiDARベースの手法は、これらのデータをLiDAR信号として扱い、ポイントワイズネットワークを使用して、これらの結果を予測する(c)。しかし、PatchNetは、その後の処理のために画像表現として整理する(d)。

具体的には、[35]は、既製のLiDARベースの3D検出器[28,18]を採用し、生成された擬似LiDAR信号を直接処理する。AM3D[24]は、生成された擬似LiDAR表現に補完的なRGBキューを埋め込むためのマルチモーダル特徴融合モジュールを提案する。また、[24]では、点群アノテーションの不正確さに起因する問題を回避するために、深度事前分布に基づく背景点セグメンテーションモジュールを提案している。[36]は、局所的なズレの問題を緩和することができる2D-3Dバウンディングボックス整合性損失を提案している。しかし、このような方法は深度マップの精度に大きく依存する。全体として、擬似LiDARベースの検出器は3D検出タスクにおいて素晴らしい精度を達成しているが、その基礎となるメカニズムは研究コミュニティにとってまだ不明瞭である。本論文では、この問題に関して詳細な調査を行う。また、擬似LiDARベースの検出器は、生成された3Dデータを点群として扱い、点群の処理にPointNetを使用するが、我々のPatchNetはそれらを画像として整理し、データの処理に2D CNNを使用することを容易にする。

3 擬似LiDAR表現への変換

本節では、擬似LiDAR表現が3次元検出精度に与える影響について検討する。特に、まず擬似LiDARベースの検出器について簡単にレビューし、その画像ベースの等価検出器の技術的詳細を紹介する。

次に、これら2つの検出器の性能を比較することで、データ表現が性能向上の内部的な理由であるかどうかを分析する。

3.1 擬似LiDARベースの検出器のレビュー

ここでは、擬似LiDAR[35]を例として解析を行い、[35]のパラダイムをまとめると以下のようになる：

Step 1: Depth estimation. Given a single monocular image (or stereo pairs) as input, [35] predict the depth d for each image pixel (u, v) using a stand alone CNN (Fig 1(a)).

Step 2: 2D detection. Another CNN is adopted to generate 2D object region proposals (Fig 1(a)).

ステップ3:3Dデータの生成。まず、ステップ2で生成された領域提案に従って、ステップ1で生成された深度マップから関心領域(ROI)を切り出す。次に、各ROIの画素の3次元座標を以下のようにして復元する：

$$\begin{cases} z = d, \\ x = (u - C_x) \times z/f, \\ y = (v - C_y) \times z/f, \end{cases} \tag{1}$$

ここで、fはカメラの焦点距離、(C_x, C_y)は主点である(図1(b))。ステップ4:3Dオブジェクト検出。擬似LiDARベースのアプローチは、ステップ3で生成された3Dデータを LiDAR信号として扱い、ポイントワイズCNNを用いてそこから結果を予測する(図1(c))。特に、$x_i \in_R^d$ を持つ非順序点集合 {x_1, x_2, ・・・, x_n } として扱い、点集合を出力ベクトルに写像する集合関数 f を定義する PointNet で処理する：

$$f(x_1, x_2, ..., x_n) = \gamma \left(\underset{i=1,...,n}{\mathbf{MAX}} \{h(x_i)\} \right) \tag{2}$$

where γ and h are implemented by multi-layer perceptron (MLP) layers.

3.2 PatchNet-vanilla:擬似LiDARの等価実装

分析。擬似LiDARに基づくアプローチ[24,35]と他のアプローチとの最も大きな違いは、深度マップの表現にある。24,35]の著者は、擬似LiDAR表現が物体の3次元構造を記述するのに適しており、これが彼らのモデルの精度が高い主な理由であると主張している。これを検証するために、入力表現以外は擬似LiDAR[35]と同じである画像表現ベースの検出器、すなわちPatchNet-vanillaを実施する。

実装。PatchNet-vanillaのステップ1、2、3は、擬似LiDARベースの検出器と同じである。したがって、推定奥行き、2次元検出結果、生成された3次元データは同じである。主な違いはステップ4であり、これについては詳細に分析する。具体的には、PatchNet-vanillaでは

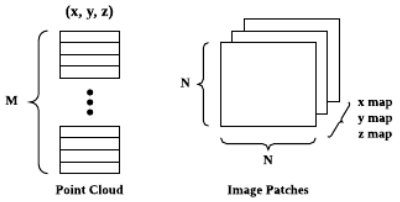

図2. 入力データの説明図擬似LiDARベースのアプローチは点群(左)を入力とし、PatchNetは画像パッチ(右)を入力とする。この2種類の入力データが同じ量の情報を含むように、M = N ×N とする。

生成された3次元データは画像表現として構成され(図2参照)、各画素位置は3チャンネル、すなわち式1の(x, y, z)である。擬似LiDARで用いられるポイントワイズCNNとは異なり、PatchNet-vanillaでは入力データの処理に2次元CNNが用いられている(図1(d))。なお、1×1の受容野とグローバルマックスプーリングによる2次元畳み込みで、式2と同じ関数を定義することができる。この方式はPointNetの公式実装[1]でも採用されている。

3.3 予備的結論

表1. 異なる入力表現の比較。実験はKITTI検証セットで行った。* は、本手法が当社で再現されたことを示す。指標は自動車カテゴリーのAP $|_{R11}$ である。

Method	Modality	3D detection			BEV detection	
		中程度の硬さの容易さの中程度の硬さ				
擬似LiDAR [35] 擬似LiDAR 28.2 擬似LiDAR		18.5	16.4	40.6	26.3	22.9
* 擬似LiDAR 28.9 パッチネットバニラ画像		18.4	16.2	41.0	26.2	22.8
28.7		18.4	16.4	40.8	26.1	22.8

PatchNet-vanillaとpseudo-LiDARの性能をTab. 1では、実装の詳細による影響を排除するために、擬似LiDARを再現している。このように、PatchNet-vanillaはpseudo-LiDARとほぼ同じ精度を達成しており、データ表現の選択は3D検出タスクに大きな影響を与えないことがわかる。さらに、データ内容に対するアブレーション研究を行い、座標変換が性能向上の重要な要因であることを観察した(実験結果と分析は5.2節に記載)。

以上の観察から、擬似LiDAR表現は必要なく、生成された3D情報を統合した後、画像表現も同じ可能性を持つことが明らかになった。さらに重要なことは、ポイントワイズCNN[29,30]と比較して、画像ベースの表現はよく研究された2次元CNNを利用することができることである

[1] https://github.com/charlesq34/pointnet.

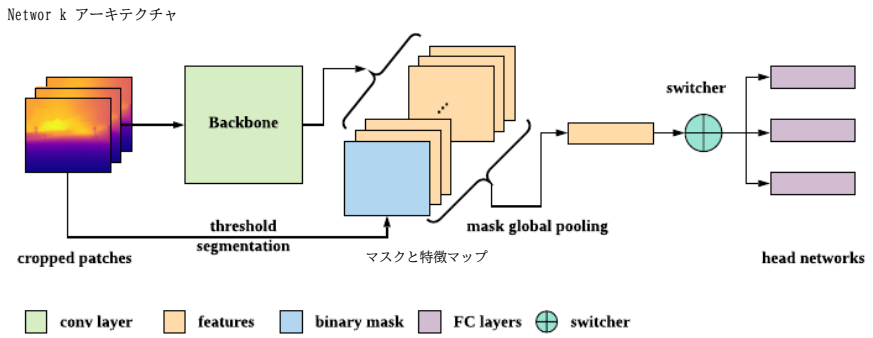

図3. ネットワークアーキテクチャの説明図x, y, z}チャンネルを持つ入力パッチが与えられたとき、まず平均深度に従ってバイナリマスクを生成し、それを用いて前景オブジェクトに対応する特徴を抽出するためのプーリング層をガイドする。次に、それぞれのヘッドネットワークの予測難易度に応じて、例を割り当てる。

高性能な3D検出器を開発するためのこの方向に沿って、提案するPatchNetフレームワークが、Sec.4で検出性能をさらに向上させるためにどのように使用されるかを示す。

4 PatchNet

PatchNetでは、まず2つの中間予測タスク(すなわち、2D検出と奥行き推定)で2つのディープCNNを訓練し、PatchNet-バニラや擬似LiDARベースの検出器と同じ位置情報と奥行き情報を得る(図1(a))。次に、図3に示すように、検出された各2次元物体提案について、深度マップから対応する領域を切り出し、式1を用いてその空間情報を復元する。次に、ROIの深層特徴をバックボーンネットワークで抽出し、マスクグローバルプーリングと前景マスクでフィルタリングする。最後に、難易度割り当て機構を持つ検出ヘッドを用いて、(x, y, z, h, w, l, θ)でパラメータ化された3次元バウンディングボックスを予測する。

バックボーン既存のバックボーンネットワークの多くは、本手法で画像特徴を抽出するために使用することができる。我々の実装では、3D検出のバックボーンとして、Squeeze-and-Excitation (SE) ブロック [16] を持つResNet-18 [15] を使用している。さらに、元のSE-ResNet-18の出力特徴量が入力画像パッチと同じサイズになるように、全てのプーリング層を削除する。次に、マスクグローバルプーリング演算と生成マスクを用いて、前景オブジェクトから特徴を抽出する。

マスクグローバルプーリングバックボーンネットワークから出力された特徴マップXは、グローバルプーリングにより特徴ベクトルに変換される。

従来のグローバルプーリングは、すべての位置の特徴を考慮し、グローバルな特徴を出力する。よりロバストな特徴を得るために、前景領域内の特徴のみに対してグローバルプーリングを行い、最終的な特徴が関心のある画素に対応するようにする。具体的には、前景領域を示すバイナリマスクMを追加で生成する。このマスクは、グローバルプーリングの前に、前景特徴を選択するために特徴マップXに適用される。このようなマスクグローバルプーリングは、最終的な特徴が関心領域に集中することを促す。

マスク生成先行研究[24]に従い、深度マップに閾値を設定することで、前景/後景の2値マスクMを得る。具体的には、経験的に各パッチの平均深度にオフセットを追加し、閾値とする。この閾値より小さい深度値を持つ領域を前景領域とみなすことにする。2値マスクMは入力画像と同じ解像度を持ち、その値は前景領域に対応するものを1、それ以外を0とする。

Head KITTIデータセットで採用されている難易度別評価に触発され、異なる難易度のサンプルを個別に扱うために3つのブランチを使用する。分岐を選択するには、特定のモジュールが必要である。具体的には、3つの並列ボックス推定器に特徴マップを送信する前に、各インスタンスの難易度を予測する別のブランチを追加する。

3つの分岐はネットワークアーキテクチャですべて同じであり、異なる難易度を扱うための学習パラメータでのみ異なることに注意。また、我々の実装では、3つのブランチとも同時に結果を予測し、そのうちの2つは難易度予測器の出力に従ってブロックされる。理論的には、これはアルゴリズムの精度に影響を与えず、余分なGPUメモリのコストですべてのブランチを並列に実行することができる。

損失関数(Loss function) グラウンドトゥルースボックスは、中心(x, y, z)、サイズ(w, h, l)、方位角θでパラメータ化される。ベースラインモデルには、[28]で提案された損失関数を採用した:

$$\mathcal{L} = \mathcal{L}_{center} + \mathcal{L}_{size} + \mathcal{L}_{heading} + \lambda \cdot \mathcal{L}_{corner} \tag{3}$$

ここで、L_{center} 、L_{size} 、$L_{heading}$ はそれぞれ中心、大きさ、方位角の損失関数を表す。λは経験的重みであり、L_{corner}は潜在的な最適でない問題を緩和するために用いられる。詳細は[28]を参照されたい。

5 Experiments

5.1 Setup

データセット KITTIデータセット[11]を用いて、本アプローチを評価する。このデータセットは、学習用に7,481枚、テスト用に7,518枚の画像を提供する。検出と

ローカライゼーション(鳥瞰検出)タスクは、オブジェクトのオクルージョンと切り捨てレベルに応じて、簡単、中程度、難しいの3つの異なるサブセットで評価される。テストセットのグランドトゥルースは利用できず、テストサーバーへのアクセスも限られているため、先行研究[6,7,8]のプロトコルに従い、学習データを学習セット(3,712画像)と検証セット(3,769画像)に分割する。この分割に基づいてアブレーション研究を行い、KITTIサーバーが提供するテストセットでの最終結果も報告する。紙面の都合上、本論文では単眼画像のCar検出結果のみを報告する。ステレオペアと歩行者/自転車に関する詳細は付録を参照されたい。

Metric 先行研究の多くは、以下のように11点補間平均精度(IAP)メトリック[11]を用いている。

$$AP|_{R_{11}} = \frac{1}{11} \sum_{r \in R_{11}} \underset{\tilde{r} \geq r}{max}\, \rho(\tilde{r}). \tag{4}$$

最近、KITTIと[34]は、性能の向上を避けるために、"0 "を除いた新しい40点IAP（AP $|_{R40}$ ）と、Precision/Recall曲線下の面積をよりよく近似するための4倍密度の補間予測を求めている。以下の実験では、先行研究および将来の研究との公平かつ包括的な比較のために、AP $|_{R11}$ と AP $|_{R40}$ の両方を示す。

5.2 擬似LiDAR表現の検討

表2. KITTI検証セットにおける3次元物体検出結果。指標は、11の想起位置を持つCarカテゴリのAP_{3D}とAP_{BEV}である。* は、メソッドが私たち自身によって再現されていることを示す。

Method	Modality	3D detection			BEV detection		
		中程度の硬さの容易さの中程度の硬さ					
pseudo-LiDAR [35]	pseudo-LiDAR	28.2	18.5	16.4	40.6	26.3	22.9
pseudo-LiDAR*	pseudo-LiDAR	28.9	18.4	16.2	41.0	26.2	22.8
AM3D [24]	pseudo-LiDAR	32.2	21.1	17.3	43.8	28.4	23.9
PatchNet-vanilla	image	28.7	18.4	16.4	40.8	26.1	22.8
PatchNet-AM3D	image	32.8	20.9	17.3	43.5	28.2	23.6
PatchNet	image	35.1	22.0	19.6	44.4	29.1	24.1
Improvement	-	+2.9	+0.9	+2.3	+0.6	+0.7	+0.2

データ表現の解析 Tab.2に示すように、PatchNet-vanillaは、擬似LiDARと同等の結果を示している。2に示すように、PatchNet-vanillaは擬似LiDARと同等の結果を示しており、データ表現が3D検出器の性能を向上させる重要な要素ではないことを示している。この主張をさらに検証するために、AM3Dに基づく画像表現ベースの検出器も調整し、再び一致した性能を達成した。

表3. KITTI検証セットにおける異なる入力データの比較。指標は、llの想起位置を持つCarカテゴリのAP_{3D}とAP_{BEV}である。

input	AP_{3D}			AP_{BEV}		
	中程度の硬さの容易さの中程度の硬さ					
$\{z\}$	4.51	3.48	3.03	6.31	4.50	3.98
$\{x, z\}$	27.1	18.3	15.8	35.9	23.4	18.3
$\{x, y, z\}$	35.1	22.0	19.6	44.4	29.1	24.1
$\{u, v, z\}$	24.6	15.7	14.6	33.2	21.3	16.7

データ内容の分析入力チャンネルの効果に関するアブレーション研究を行い、その結果をTab. 5.2. この結果から、奥行きのみを入力として使用した場合、正確な3Dバウンディングボックスを得ることはほとんど不可能であることがわかる。他の座標を使用した場合、予測されたボックスの精度が大幅に向上し、生成された空間特徴の重要性が検証された。y軸のデータがない場合、この検出精度は我々のフルモデルよりもはるかに悪いことに注意すべきである。これは、すべての座標が3D検出に有用であることを示している。
擬似LiDARでは、画像の座標(u, v)はカメラ情報を用いてワールド座標(x, y)に投影される。表5.2の実験結果も比較した。5.2の実験結果も、異なる座標系の有効性を比較したものである。実験結果によると、カメラ情報を利用するワールド座標(x, y)は、画像座標(u, v)よりもはるかに良い性能を示す。以上の実験を通して、データ表現そのものではなく、座標系変換が本当に重要であることが観察できる。

5.3 PatchNetの性能向上

バックボーン(Backbone)(擬似)LiDARベースの手法で一般的に使用されるポイントワイズバックボーンネットと比較して、[15,16,37]のような標準的な2Dバックボーンは、より識別性の高い特徴を抽出することができ、これは画像ベースの検出器の自然な利点である。提案するPatchNetに対するバックボーンの違いによる影響を調査し、実験結果をTab. 4(左)にまとめた。オリジナルのPointNetは8層しかない。公正な比較のために、18層からなるPointNetを構築し、これをTab.4のPointNet-18と表記する。4. PointNet-18と比較すると

2次元畳み込みバックボーンは、特にハードセッティングにおいて、3次元ボックスの精度を向上させることができる。これは、これらのケースは通常、オクルード/切り捨て、またはカメラから遠く離れており、その姿勢を推定することは、コンテキスト情報に依存するためである。しかし、ポイントワイズCNNは、データの局所的な特徴を効率的に抽出することが困難であることがわかる。このような観点から、画像表現に基づく検出器は、より大きな開発ポテンシャルを持つ。その上、Tab. 4(右)から、CNNがResNeXt-18からResNeXt-50まで多くの層を持つ場合、精度はあまり向上しないことがわかる。ResNeXt-50と比較して、ResNeXt-101は性能が悪く、これはオーバーフィッティングに起因する。すべてのCNNはゼロから学習される。

表4. KITTI検証セットにおける異なるバックボーンネットの比較。メトリクスは$AP_{3D}|_{R11}$で、IoU閾値=0.7のCarカテゴリの3D検出タスクである。その他の設定はPatchNet-vanillaと同じ。

Backbone	Easy	Moderate	Hard
PointNet-18	31.1	20.5	17.0
ResNet-18	33.2	21.3	19.1
ResNeXt-18	33.4	21.2	19.2
SE-ResNet-18	33.7	21.5	19.2

Backbone	Easy	Moderate	Hard
ResNeXt-18	32.7	21.2	19.2
ResNeXt-50	32.9	21.4	17.3
ResNeXt-101	31.1	20.9	17.0

マスクグローバルプーリング。PatchNetでは、特徴マップを注目画素の集合から抽出するようにマスクグローバルプーリング操作を設計しており、これはハードアテンションメカニズムとみなすことができる。表? 例えば、マスクグローバルプーリング(max)は、$AP_{3D}|_{11}$ を中程度の設定で1.4%、簡単な設定で2.7%改善することができ、maxプーリングはavgプーリングよりわずかに優れている。また、図?に示す可視化結果も参考までにご確認ください。は、性能向上の理由を直感的に説明することができる。具体的には、マスクグローバルプーリングによってフィルタリングされた活性化ユニットのほとんどは前景ゴールに対応し、標準的なグローバルマックスプーリングからのものは背景に多くの活性化ユニットを持つことになる。
背景点は本モデルでは文脈情報を提供するが、PointNetの入力として[28,35]には関与していないことに注意する必要がある。

インスタンスの割り当て。スタンド単独モジュールを使って各インスタンスの「難易度」を予測し、対応するヘッドネットワークに割り当てる。表 5.3 にこの機構のアブレーションスタディを示す。まず、出力の精度はインスタンスの割り当てによって向上することがわかる。興味深いことに、すべてのケースで「難易度」のアノテーションが得られるわけではないことを考慮し、単純な代替案を使用する:オブジェクトの「難易度」を表現するためにオブジェクトからカメラまでの距離を使用し(我々のデフォルト設定)、この実験で使用される閾値は(30, 50)である。実験によると、この方式は予測された難易度と同程度の性能を得ることができた。

表5. KITTI検証セットにおけるマスクグローバルプーリングのアブレーション研究。指標は、11の想起位置を持つCarカテゴリのAP_{3D}とAP_{BEV}である。その他の設定はPatchNet(フルモデル)と同じ。

pooling	type	AP_{3D}			AP_{BEV}		
		中程度の硬さの容易さの中程度の硬さ					
standard	max	32.4	20.6	17.7	41.3	27.0	21.6
mask	avg	34.6	21.6	19.3	43.5	28.7	23.3
mask	max	35.1	22.0	19.6	44.4	29.1	24.1

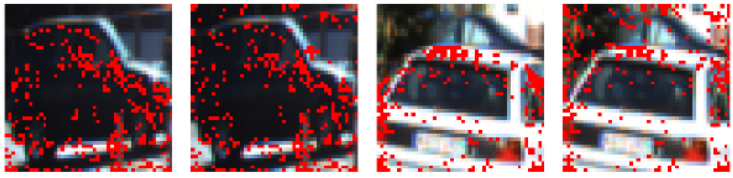

図4. KITTI検証セットにおける最大グローバルプーリングの定性的比較。各画像ペアの左/右画像は、マスク/標準グローバルプーリングによって活性化されたユニットを示す。

表6. KITTI検証セットにおけるインスタンス割り当てのアブレーション研究。指標は、11の想起位置を持つCarカテゴリのAP_{3D}とAP_{BEV}である。

assignment	switcher	AP_{3D}			AP_{BEV}		
		中程度の硬さの容易さの中程度の硬さ					
-	-	33.7	21.5	19.2	42.5	28.2	23.5
✓	difficulty	34.7	22.1	19.5	44.1	29.0	24.2
✓	distance	35.1	22.0	19.6	44.4	29.1	24.1

5.4 最先端手法との比較

表7に示すように 7に示すように、KITTIデータセットにおける自動車カテゴリの3D検出結果を報告する。提案するPatchNetは、公開されている全ての手法の中で1位である(適度な設定でランク付け)。全体として、我々の手法は、簡単なレベルのテストセットを除く全ての設定において、他の最先端手法よりも優れた結果を達成している。例えば、KITTIデータセットで最も困難な3つのメトリクスにおいて、ハード設定下で現在の最先端AM3D [24]を0.65/1.56/2.34上回った。また、提案手法は既存の擬似LiDARベースのアプローチを凌駕する。なお、[24,35,39,3]と同じ深度推定器(DORN)を使用しており、提案手法のパイプラインは擬似LiDARベースの対応するもの[3,39]よりもはるかに単純である。

表7. KITTIデータセットにおけるCarカテゴリの3D検出性能。テストセットでは、AP $|_{R40}$ だけが公式リーダーボードから提供される。検証セットについては、AP $|_{R40}$ と AP $|_{R11}$ の両方を報告する。IoUの閾値は0.7に設定されている。* は擬似LiDARデータに基づく方法であることを示す。方法は適度な設定(KITTIリーダーボードと同じ)でランク付けされている。最良の結果を太字で強調する。

Method	testing ($AP|_{40}$)			validation($AP|_{40}$)			validation ($AP|_{11}$)		
	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
OFTNet [33]	1.61	1.32	1.00	-	-	-	4.07	3.27	3.29
FQNet [23]	2.77	1.51	1.01	-	-	-	5.98	5.50	4.75
ROI-10D [25]	4.32	2.02	1.46	-	-	-	10.25	6.39	6.18
GS3D [19]	4.47	2.90	2.47	-	-	-	13.46	10.97	10.38
Shift R-CNN [27]	6.88	3.87	2.83	-	-	-	13.84	11.29	11.08
Multi-Fusion [38]	7.08	5.18	4.68	-	-	-	22.03	13.63	11.60
MonoGRNet [31]	9.61	5.74	4.25	-	-	-	13.88	10.19	7.62
Decoupled-3D* [3]	11.08	7.02	5.63	-	-	-	26.95	18.68	15.82
MonoPSR [18]	10.76	7.25	5.85	-	-	-	12.75	11.48	8.59
MonoPL* [36]	10.76	7.50	6.10	-	-	-	31.5	21.00	17.50
SS3D [17]	10.78	7.68	6.51	-	-	-	14.52	13.15	11.85
MonoDIS [34]	10.37	7.94	6.40	11.06	7.60	6.37	18.05	14.98	13.42
M3D-RPN [2]	14.76	9.71	7.42	-	-	-	20.27	17.06	15.21
PL-AVOD* [35]	-	-	-	-	-	-	19.5	17.2	16.2
PL-FPointNet* [35]	-	-	-	-	-	-	28.2	18.5	16.4
AM3D* [24]	**16.50**	10.74	9.52	28.31	15.76	12.24	32.23	21.09	17.26
PatchNet	15.68	**11.12**	**10.17**	**31.6**	**16.8**	**13.8**	**35.1**	**22.0**	**19.6**

これは、我々の設計の有効性を示している。また、提案モデルはテストセットでの簡単な設定において、AM3D [24]に遅れをとっていることがわかる。これは、2D検出器の違いに起因すると考えられる。簡単な分割は例数が最も少ないので、この設定の性能は変動しやすいことを強調する。また、これら3つの分割は封じ込め関係であることに注意してください(例えば、ハード分割はすべてのインスタンスが容易な設定と中程度の設定に属しています)。

5.5 定性的結果

図5に、我々のPatchNetモデルの代表的な出力を可視化する。妥当な距離にある単純なケースでは、我々のモデルは驚くほど正確な3Dバウンディングボックスを出力することが観察できる。相対的に、遠い物体については、その中心を決定することは難しいが、その大きさと方位角の推定はまだ正確である。

一方、いくつかの失敗パターンが観察され、今後の取り組みの方向性が示された。第一に、我々の手法はしばしば切り捨てられた/閉じたオブジェクトで間違いを犯し、しばしば不正確な方位推定値として現れる。第二に、我々の2D検出器は強いオクルージョンのために物体を見逃すことがあり、そのためこれらのサンプルは後続のステップで無視される。

図5. KITTI検証セットでの定性的結果。赤のボックスは我々の予測を表し、緑のボックスはグランドトゥルースからのものである。LiDAR信号は可視化にのみ使用される。ズームインしてカラーで見るのがベスト。

6 Conclusions

本論文では、擬似LiDAR表現に基づく3D検出器が有望な性能を達成する基本的な原因を探るために、新しいネットワークアーキテクチャ、すなわちPatchNetを提案する。他の研究とは異なり、我々は、点群表現そのものではなく、カメラパラメータによって画像座標をワールド座標に投影することが重要な要素であると主張する。さらに重要なことは、ワールド座標表現を画像表現に容易に統合できることで、より柔軟で成熟した2D CNN技術を用いた3D検出器の性能をさらに高めることができる。KITTIデータセットでの実験結果は、我々の議論を実証し、画像表現に基づく3D検出器の可能性を示している。これらの新しい視点が、単眼/ステレオ3D物体検出コミュニティに洞察を提供し、画像ベースの3D検出のための新しい2D CNN設計の開発を促進することを期待する。

7 謝辞

本研究は、SenseTime、オーストラリア研究評議会助成金DP200103223、オーストラリア医学研究未来基金MRFAI000085の支援を受けた。

References

1. Alhashim, I., Wonka, P.: High quality monocular depth estimation via transfer learning. arXiv e-prints **abs/1812.11941**, arXiv:1812.11941 (2018), `https://arxiv.org/abs/1812.11941`
2. Brazil, G., Liu, X.: M3d-rpn: Monocular 3d region proposal network for object detection. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
3. Cai, Y., Li, B., Jiao, Z., Li, H., Zeng, X., Wang, X.: Monocular 3d object detection with decoupled structured polygon estimation and height-guided depth estimation. arXiv preprint arXiv:2002.01619 (2020)
4. Chabot, F., Chaouch, M., Rabarisoa, J., Teuliere, C., Chateau, T.: Deep manta: A coarse-to-fine many-task network for joint 2d and 3d vehicle analysis from monocular image. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2040–2049 (2017)
5. Chang, J.R., Chen, Y.S.: Pyramid stereo matching network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5410–5418 (2018)
6. Chen, X., Kundu, K., Zhang, Z., Ma, H., Fidler, S., Urtasun, R.: Monocular 3d object detection for autonomous driving. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2147–2156 (2016)
7. Chen, X., Kundu, K., Zhu, Y., Berneshawi, A.G., Ma, H., Fidler, S., Urtasun, R.: 3d object proposals for accurate object class detection. In: Advances in Neural Information Processing Systems. pp. 424–432 (2015)
8. Chen, X., Ma, H., Wan, J., Li, B., Xia, T.: Multi-view 3d object detection network for autonomous driving. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
9. Dai, J., Li, Y., He, K., Sun, J.: R-fcn: Object detection via region-based fully convolutional networks. In: Advances in neural information processing systems. pp. 379–387 (2016)
10. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2002–2011 (2018)
11. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 3354–3361. IEEE (2012)
12. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 1440–1448 (2015)
13. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 270–279 (2017)
14. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
16. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
17. Jrgensen, E., Zach, C., Kahl, F.: Monocular 3d object detection and box fitting trained end-to-end using intersection-over-union loss. CoRR **abs/1906.08070** (2019), `http://arxiv.org/abs/1906.08070`

18. Ku, J., Mozifian, M., Lee, J., Harakeh, A., Waslander, S.L.: Joint 3d proposal generation and object detection from view aggregation. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1–8. IEEE (2018)
19. Li, B., Ouyang, W., Sheng, L., Zeng, X., Wang, X.: Gs3d: An efficient 3d object detection framework for autonomous driving. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1019–1028 (2019)
20. Li, P., Chen, X., Shen, S.: Stereo r-cnn based 3d object detection for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
21. Lin, T.Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
22. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
23. Liu, L., Lu, J., Xu, C., Tian, Q., Zhou, J.: Deep fitting degree scoring network for monocular 3d object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1057–1066 (2019)
24. Ma, X., Wang, Z., Li, H., Zhang, P., Ouyang, W., Fan, X.: Accurate monocular 3d object detection via color-embedded 3d reconstruction for autonomous driving. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
25. Manhardt, F., Kehl, W., Gaidon, A.: Roi-10d: Monocular lifting of 2d detection to 6d pose and metric shape. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
26. Mousavian, A., Anguelov, D., Flynn, J., Kosecka, J.: 3d bounding box estimation using deep learning and geometry. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7074–7082 (2017)
27. Naiden, A., Paunescu, V., Kim, G., Jeon, B., Leordeanu, M.: Shift r-cnn: Deep monocular 3d object detection with closed-form geometric constraints. In: 2019 IEEE International Conference on Image Processing (ICIP). pp. 61–65. IEEE (2019)
28. Qi, C.R., Liu, W., Wu, C., Su, H., Guibas, L.J.: Frustum pointnets for 3d object detection from rgb-d data. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
29. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 652–660 (2017)
30. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In: Advances in neural information processing systems. pp. 5099–5108 (2017)
31. Qin, Z., Wang, J., Lu, Y.: Monogrnet: A geometric reasoning network for monocular 3d object localization. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 8851–8858 (2019)
32. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. pp. 91–99 (2015)
33. Roddick, T., Kendall, A., Cipolla, R.: Orthographic feature transform for monocular 3d object detection. arXiv preprint arXiv:1811.08188 (2018)

34. Simonelli, A., Bulo, S.R., Porzi, L., Lopez-Antequera, M., Kontschieder, P.: Disentangling monocular 3d object detection. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
35. Wang, Y., Chao, W.L., Garg, D., Hariharan, B., Campbell, M., Weinberger, K.Q.: Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
36. Weng, X., Kitani, K.: Monocular 3d object detection with pseudo-lidar point cloud. In: IEEE International Conference on Computer Vision (ICCV) Workshops (Oct 2019)
37. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1492–1500 (2017)
38. Xu, B., Chen, Z.: Multi-level fusion based 3d object detection from monocular images. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
39. You, Y., Wang, Y., Chao, W.L., Garg, D., Pleiss, G., Hariharan, B., Campbell, M., Weinberger, K.Q.: Pseudo-lidar++: Accurate depth for 3d object detection in autonomous driving. arXiv preprint arXiv:1906.06310 (2019)
40. Zhou, D., Zhou, X., Zhang, H., Yi, S., Ouyang, W.: Cheaper pre-training lunch: An efficient paradigm for object detection. arXiv preprint arXiv:2004.12178 (2020)

A Overview

この文書は、本論文に追加の分析と追加の実験を提供する。具体的には、Sec.Bにおいて、提案手法のレイテンシを分析し、いくつかの擬似LiDARベースの手法と比較する。第C節ではステレオ画像の性能を示し、第D節では歩行者検出とサイクリスト検出の結果を示す。最後に、Sec.Eでは、より多くの可視化例を示す。

B ランタイム解析

本節では、我々のPatchNetのレイテンシを分析し、擬似LiDAR表現に基づくいくつかの既存手法[24,35,36]と比較する。一般に、4つの方法はすべて3つの主要な段階に分けられる。我々の設計では、PatchNet-vanillaとpseudo-LiDARの処理フローは同じであるが、入力の表現が異なる。つまり、これら2つの手法の実行時間はほぼ同じであり、以下のように示される(1080GPUでテスト):

表8. PachNet-vanillaと擬似LiDARの実行時間。

2次元検出	奥行き推定	3次元検出
60ms	400ms	28ms

PatchNetは同じ2D検出器と奥行き推定器を共有しており(異なる奥行き推定器の実行時間は大きく異なることに注意、詳細はKITTIベンチマークを参照)、異なるバックボーンモデルに対する3D検出ステージの実行時間を以下のように示す:

表9. 3D検出段階におけるPatchNetの実行時間。

バックボーン	PointNet-18	ResNet-18	ResNeXt-18	SE-ResNet-18
runtime	12ms	23ms	18ms	26ms

PatchNetにいくつかの追加操作を加えたが、ベースラインモデル(PointNet-18)の実行時間は12ms、PatchNet-vanillaの実行時間は28msである。これは主に、前景セグメンテーションネットを削除し、前景セグメンテーションに動的閾値を使用するため、約18msを節約できるためである。最良のバックボーンでは、実行時間はわずか26msであり、3d検出のための擬似LiDARの実行時間は同程度である。

また、PatchNetと[24]は同じセグメンテーション手法を用いているが、[24]は画像特徴を抽出するために別のResNet-34を追加している。36]では、2Dインスタンスセグメンテーションネットを追加しており、多くの計算オーバーヘッド(例えば、約200msのMask RCNN [14])をもたらす。
要約すると、PatchNetは[24,36]よりも効率的であり、[35]と同様の実行時間を持つ。

C ステレオ画像

擬似LiDAR表現は、ステレオ3D検出タスクの分野でも広く使われている。提案手法が両眼画像でも有効であることを検証するため、単眼深度マップをステレオ深度マップに置き換え(ステレオ深度推定器としてPSMNet[5]を使用し、[35]から事前学習済みモデルを取得)、AP $|_{R11}$ を用いてKITTI検証セットで性能をテストし、先行研究との比較をより良くする。表10に示すように 10 に示すように、PatchNetvanilla は pseudo-LiDAR とほぼ同じ精度であるが、PatchNet はより良い性能を達成している。参考までにAP $|_{R40}$も報告する。

表10. KITTI検証データセットにおけるCarカテゴリのステレオ3D検出性能。IoUの閾値は0.7に設定されている。最良の結果を太字で強調する。

Method	3D Detection			BEV Detection		
	中程度の硬さの容易さの中程度の硬さ					
3DOP [7]	6.55	5.07	4.10	12.63	9.49	7.59
Multi-Fusion [38]	-	9.80	-	-	19.54	-
Stereo-RCNN [20]	54.1	36.7	31.1	68.5	48.3	41.5
Pseudo-LiDAR [35]	59.4	39.8	33.5	72.8	51.8	44.0
PatchNet-vanilla	60.8	40.1	33.6	72.7	51.2	43.8
PatchNet	**65.9**	**42.5**	**38.5**	**74.5**	**52.9**	**44.8**
PatchNet-vanilla@$AP\|_{R_{40}}$	61.4	37.6	31.6	73.5	49.8	41.7
PatchNet@$AP\|_{R_{40}}$	**66.0**	**41.1**	**34.6**	**76.8**	**52.8**	**44.3**

D 歩行者と自転車

比較のため、KITTI検証セットにおける3D検出タスクの歩行者/自転車検出性能も報告する。具体的には、AP $|_{R11}$ を指標として、単眼画像とステレオ画像の両方を用いて実験を行う。表11からわかるように 11から、提案モデルも各設定において[35]よりも良い性能を得ていることがわかる。なお、擬似LiDARは単眼画像に対する歩行者/自転車検出の結果を提供していないため、公式コードを用いて評価した。

また、歩行者/自転車の検出精度は、自動車の検出精度に比べて大きく変動している。この性能の変動は、主に学習サンプルの不足が原因である(KITTI学習セットの歩行者/自転車の学習サンプルは2,207/734個しかないが、14,357個の自動車インスタンスが提供される)。この問題は、より多くの学習データを導入するか、より効果的なデータ補強戦略を導入することで軽減できる。

表11. KITTI検証データセットにおける歩行者/自転車カテゴリの3D検出性能。指標は$AP|_{R11}$、IoU閾値は0.5に設定。最良の結果を太字で示す。

Method	Category	Monocular			Stereo		
		中程度の硬さの容易さの中程度の硬さ					
Pseudo-LiDAR [35]	Pedestrian	7.32	6.19	5.64	33.8	27.4	24.0
PatchNet	Pedestrian	**9.82**	**7.86**	**6.84**	**38.8**	**30.1**	**26.5**
Pseudo-LiDAR [35]	Cyclist	5.49	3.85	3.82	41.3	25.2	24.9
PatchNet	Cyclist	**8.14**	**4.84**	**4.62**	**46.8**	**29.0**	**26.8**

E より定性的な例

このパートでは、図6にいくつかの代表的な定性的結果によって、単眼画像とステレオ画像のペアを比較する。まず、ステレオ画像は物体をより正確に検出できることがわかるが、これは一般的に、大きさや方位推定ではなく、より良い奥行き推定に反映される。そして、ほとんどの近距離物体において、視覚経験の点で、単眼画像はステレオ画像に劣らない(ただし、これらのインスタンスにはまだいくつかの失敗例がある)。

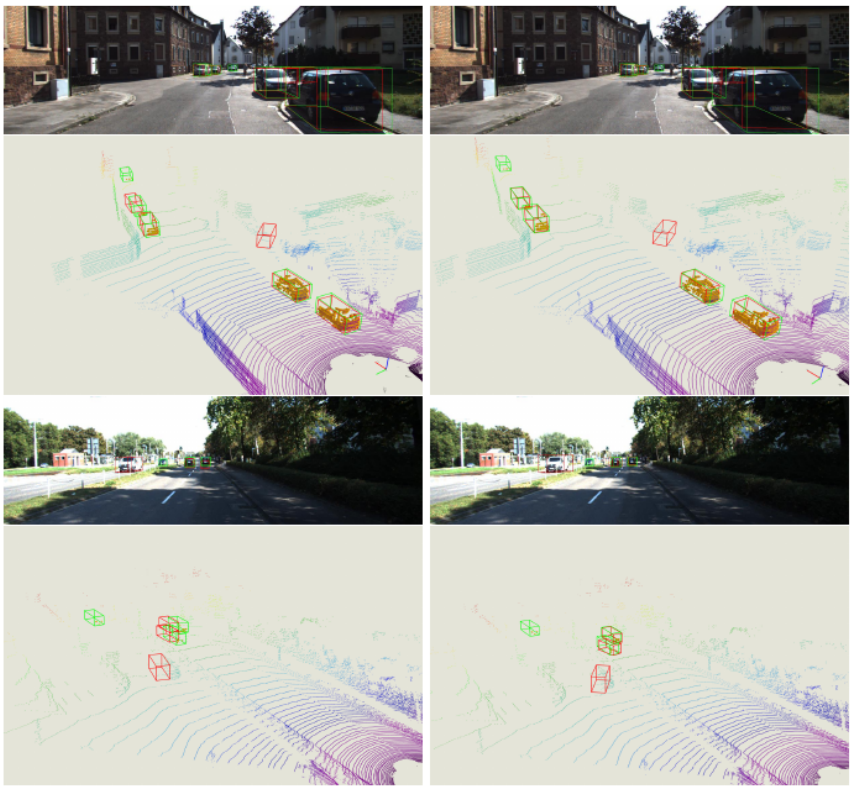

図6. KITTI検証セットでの定性的結果。左:単眼検出結果。右:ステレオ検出結果。赤いボックスは我々の予測を表し、緑のボックスはグランドトゥルースからのものである。LiDAR信号は可視化にのみ使用される。ズームインしてカラーで見るのがベスト。